View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Xiamen University Institutional Repository

学校编码：10384　　分类号_____密级_____
学号：24320071151829　　UDC _____

厦门大学

硕士学位论文

子空间聚类方法研究及其在垃圾邮件过滤中的应用

Research on Subspace Clustering Methods and their Applications in Spam Filtering

黄王非

指导教师姓名：姜青山教授

专　业　名　称：计算机软件与理论

论文提交日期：2010 年5 月

论文答辩时间：2010 年　月

学位授予日期：2010 年　月

答辩委员会主席：_____

评　　阅　　人：_____

2010 年 5 月

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

摘 要

聚类分析是数据挖掘的基础功能之一，在生物信息学、信息过滤、资料自动分类等领域得到广泛应用。然而在众多实际应用中，实际数据变得越来越庞大和复杂，并且具有非常高的维度。受“维度效应”的影响，传统的聚类算法在高维数据上的聚类精度不尽如人意。因而研究高维数据的聚类分析方法具有非常重要的意义，已成为近年来研究的一个重点课题。

在高维空间中簇类可能只存在于某些低维子空间中，而不同的簇类其所处的子空间也可能存在差异。因而子空间聚类方法成为高维数据聚类分析中的一个研究热点。局部维度约简方法和簇所处子空间的发现方法是子空间聚类方法研究领域的两个关键问题。我们从局部维度约简方法入手，通过分析现有方法的不足，改进了一种局部维度约简方法；并以此为基础，分析了对象在子空间中的局部分布密度，提出了一种基于局部密度的子空间聚类的方法；最后在垃圾邮件过滤领域进行应用研究。论文的主要工作如下：

1. 改进了一种局部维度约简方法。基于单一维度的密度进行局部维度约简，克服了之前方法基于全空间中的邻域概念进行维度约简带来的影响。实验结果表明，这种改进的局部维度约简方法能够提高聚类的精度；
2. 提出了一种基于局部密度的子空间聚类方法。方法定义了对象局部密度的度量方法，结合已有的聚类算法对对象的局部密度聚类，发现簇所在的子空间，进而在子空间中找到簇类。实验结果表明，基于局部密度的子空间聚类方法能够在帮助已有算法提高聚类的精度；
3. 在以上研究的基础上，设计并实现了一个基于子空间聚类的垃圾邮件过滤系统，通过实际应用验证算法的有效性。系统也为进一步研究子空间聚类及在垃圾邮件过滤中的应用提供了基础。

关键词：高维数据；子空间聚类；垃圾邮件过滤

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

Research on Subspace Cluster Methods and their Applications in Spam Filtering

Abstract

Cluster analysis is one of the basic functions of data mining, and it has been widely used in many fields. However, in practical applications, dimensions of data are increasing. Because of the curse of dimensionality, traditional clustering algorithms cannot reach people's expectation. Thus, research on high dimensional data clustering becomes increasingly important.

In high dimensional data, clusters usually exist in some low-dimensional subspace. Thus, subspace clustering research has become a research hot spot. Local dimension reduction methods and discovering the subspace which clusters existed are two key issues in field of subspace clustering. We propose a new local dimension reduction method by analysis problems of existing methods. And base on this method, a novel subspace clustering method is presented. The methods mentioned above are used in spam filtering. The major work of this dissertation can be summarized as follows:

1. We propose a new local dimension reduction method. This method is based on neighborhood of single dimension, and it can overcome the shortcomings of prior methods. Experimental results show that the new method improve the precision of clustering;
2. By the definition of local density, a novel subspace clustering is presented. The new method combined with the existing algorithms can achieve good results in high dimensional clustering.
3. Based on the research above, we design and implement a spam filtering system, thereby verifying the methods we proposed through the system, and assisting in future research on subspace clustering and its applications in spam filtering.

Keywords: High-dimensional Data; Subspace Clustering; Spam-Filtering

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

目　录

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

Contents

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

第一章 绪论

近年来，随着数据的爆炸式增长，数据的维度呈现出“高维”的特性，使得原有的聚类算法在高维数据空间中不再有效。因此，高维数据的聚类分析方法已经成为近年来研究的一个重点问题。论文研究高维数据聚类分析中的子空间聚类方法。在这章中将分别从数据挖掘、高维数据聚类分析、子空间聚类及其实际应用等方面阐述论文的研究背景，通过分析研究现状与存在问题，阐明了本课题的研究意义，最后对论文研究内容及结构安排进行总体概述。

1.1 研究背景及选题意义

Internet 技术的迅猛发展，使得人们获取数据变得越来越容易，能够方便的在互联网上发布数据、共享数据、搜索数据和下载数据。受益于数据库技术的迅速发展以及数据库管理系统的广泛应用，积累的数据越来越多。激增的数据背后隐藏着许多重要的信息，人们希望能够对其进行更高层次的分析，以便更好地利用这些数据[1,2]。目前的数据库系统可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。面对“数据丰富而信息缺乏”的现象，一个新的研究领域——数据挖掘（Data Mining）应运而生[1,2,3]。

数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程[1,2,3]。数据挖掘是一门交叉的学科，它与计算机学、数学、统计学、经济学、生物学、语言学等学科都有密切的联系。随着 Internet 的发展，各种类型文档的数量呈爆炸性增长。为了对这些文档进行分析、组织和总结，从而提取各种文档中所隐含的联系、规则和模式，文本挖掘应运而生，目前已成为数据挖掘的一个重要分支[1,4]。

聚类分析(Clustering Analysis)已经成为数据挖掘研究领域中一个非常活跃的研究课题[1,2,3]。作为统计学的一个分支，聚类分析已经被广泛地研究了许多年，早期研究可以追溯到上世纪 40 年代[3,5]，在发展过程中一些研究曾归入统计学和

厦门大学博硕士论文摘要库

机器学习范畴。在统计学中，聚类一般称为“聚类分析”（cluster analysis），主要集中于基于距离的聚类分析；在机器学习中，聚类称为“无指导的学习”（unsupervised learning），主要体现在聚类学习的数据对象没有类别标记，需要由聚类算法自动计算[1]。由于这种无指导特性，聚类已经广泛应用在许多领域中，包括模式识别、数据分析、图像处理、市场研究等[5]。

然而，在聚类分析的许多应用领域，实际数据变得越来越庞大和复杂，且数据的维度（特征或属性）达到几十、成百或上千维。一般认为，10-20 以上维度的数据就应被纳入高维数据的范畴[1,5]。与低维空间相比数据在高维空间中的表现有很大差异，多数传统聚类方法无法获得好的聚类效果，高维数据的聚类分析已经成为数据挖掘领域中一个具有挑战性的问题[1,5-7]。

高维数据除了具有维度数目较高、数据规模较大的特点外，还普遍存在需要进行子空间聚类的特点[4,5]。具体表现为各个类别通常只和少量的特征（比如，几百维）相关联，同时对于不同类别对应的特征集合也可能不同，甚至有很大的差异[4,5]。这就要求聚类算法不仅需要对数据集进行划分，还要给出每个簇所处的子空间。这也是传统聚类方法在面对高维数据聚类时无法获得较好聚类效果的原因之一[5]。

目前，电子邮件极大的方便了人们的生活，垃圾邮件的问题随之而来。以垃圾邮件过滤为例，电子邮件文档以向量空间模型 VSM[8] (Vector Space Model)表示后具有多达几千维的特征。这种高维性与高数据量是一般数据所不具备的。且垃圾邮件文本具有明显的子空间特征。因此，研究子空间聚类的理论和方法对于丰富高维数据聚类分析的方法，拓展高维数据聚类分析在垃圾邮件过滤中的应用具有重要的意义[5,9]。

1.2 研究现状及存在问题

数据挖掘是信息技术发展到一定阶段的必然产物，被认为是二十一世纪计算机技术最前沿的研究领域之一[1,2,3]。聚类分析作为统计学、机器学习和数据挖掘等领域的交叉学科，吸引了众多研究者投身其中，使之成为数据挖掘研究领域的一个非常活跃的研究课题[2,3]。对于聚类的研究始于 20 世纪 40 年代，迄今为止国内外的研究者们提出了很多聚类算法，主要的聚类方法可以分为如下五类：基

厦门大学博硕士论文摘要库

Degree papers are in the “Xiamen University Electronic Theses and Dissertations Database”. Full texts are available in the following ways:
1. If your library is a CALIS member libraries, please log on http://etd.calis.edu.cn/ and submit requests online, or consult the interlibrary loan department in your library.
2. For users of non-CALIS member libraries, please mail to etd@xmu.edu.cn for delivery details.

厦门大学博硕士论文摘要库